DOCUMENT RESUME

ED 448 175 TM 032 136

AUTHOR Barnette, J. Jackson; McLean, James E.
TITLE The Corrected Eta-Squared Coefficient: A Value Added Approach.
PUB DATE 2000-11-00
NOTE 24p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (28th, Bowling Green, KY, November 15-17, 2000).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; Monte Carlo Methods; *Sampling; *Statistical Bias
IDENTIFIERS *Value Added Model

ABSTRACT

Eta-Squared (ES) is often used as a measure of strength of association of an effect, a measure often associated with effect size. It is also considered the proportion of total variance accounted for by an independent variable. It is simple to compute and interpret. However, it has one critical weakness cited by several authors (C. Huberty, 1994; P. Snyder and S. Lawson, 1993; and T. Snijders, 1996), and that is a sampling bias that leads to an inflated judgment of true effect. The purpose of this study was to determine the degree of inflation by determining how large ES is likely to be by chance, find methods of predicting the mean inflation, and then proposing the use of a corrected ES coefficient that is the observed ES minus the mean expected ES, a value added approach. A Monte Carlo study was set up using a number of samples from 2 to 10 and sample sizes from 5 to 100 in steps of 5. In each number of samples and sample size configuration, 10,000 one-way analysis of variance replications, using samples drawn from the unit normal distribution, were conducted for a total of 1,800,000 replications. Patterns of observed ES values were examined for influences of number and size samples. It was clear that ES was influenced by both of these factors. Trend analysis was conducted to determine equations that could be used to predict the mean chance-based ES for given number and size of samples. In a given research situation, the expected ES coefficient may be determined for comparison with the observed ES. Such an approach removes the bias cited as the major weakness of the use of ES as a measure of strength of association and makes it a more useful measure of non-chance influence. (Contains 6 figures, 3 tables, and 43 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

ED 448 175

The Corrected Eta-Squared Coefficient: A Value Added Approach

J. Jackson Barnette
University of Iowa

James E. McLean
East Tennessee State University

Paper presented at the annual meeting
Of the
Mid-South Educational Research Association
Bowling Green, Kentucky
November 2000

TM032136

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

J. Barnette

BEST COPY AVAILABLE

Abstract

Eta-Squared (ES) is often used as a measure of strength of association of an effect, a measure often associated with effect size. It is also considered the proportion of total variance accounted for by an independent variable. It is simple to compute and interpret. However, it has one critical weakness cited by several authors (Huberty, Snyder & Lawson, and Snijders) and that is a sampling bias that leads to an inflated judgment of true effect. The purpose of this research is to determine the degree of inflation by determining how large ES is likely to be by chance, find methods of predicting the mean inflation, and then proposing the use of a corrected ES coefficient which is the observed ES minus the mean expected ES, a value added approach.

A Monte Carlo study was set up using number of samples from 2 to 10 and sample sizes from 5 to 100 in steps of 5. In each number of samples and sample size configuration, 10000 one-way ANOVA replications, using samples drawn from the unit normal distribution, were conducted for a total of 1,800,000 replications.

Patterns of observed ES values were examined for influences of number and size of samples. It was clear that ES was influenced by both of these factors. Trend analysis was conducted to determine equations that could be used to predict the mean chance-based ES for given number and size of samples. In a given research situation, the expected ES coefficient may be determined for comparison with the observed ES. Such an approach removes the bias cited as the major weakness of the use of Eta-squared as a measure of strength of association and makes it a more useful measure of non-chance influence.

The Corrected Eta-Squared Coefficient:
A Value Added Approach

Eta-Squared (ES) is probably the most used measure of effect size in conjunction with ANOVA. It is a measure of the strength of association of an effect, a measure often associated with effect size. It is also considered the proportion of total variance accounted for by an independent variable. It is simple to compute and interpret. However, it has one critical weakness cited by several authors (Huberty, 1994; Snyder & Lawson, 1993; Snijders, 1996) and that is a sampling bias that leads to an inflated judgment of true effect. The purpose of this research is to determine the degree of inflation by determining how large ES is likely to be by chance, find methods of predicting the mean inflation, and then proposing the use of a corrected ES coefficient which is the observed ES minus the mean expected ES, a value added approach.

Background

The concept of effect size has been around for many years. Cohen (1969) is generally credited with coining the term. However, the development of meta-analysis by Glass, Rosenthal and others in the 1970s (e.g., Glass, 1976; 1978; Glass & Hakstian, 1969; Rosenthal, 1976, 1978) and the popularity of a book on meta-analysis in 1981 (Glass, McGaw, & Smith) are the catalysts for the interest in the concept. Numerous publications followed on applications of effect size methodology (e.g., Lynch, 1987; McLean, 1983), methods for estimating effect size and its properties (e.g., Fowler, 1988; Gibbons, Hedeker, & Davis, 1993;Hedges, 1981, 1984; Huynh, 1989; Kraemer, 1983; Reichhardt & Gollob, 1987; Thomas, 1986), extracting effect size estimates from existing studies (e.g., Hedges, 1982; Snyder & Lawson, 1993), and correcting effect size estimates (Snyder & Lawson, 1993). Another book by Wolf (1986) presented a general methodology for conducting meta-analysis including the extraction and testing of effect sizes.

Perhaps no one has had a greater impact on the use of effect sizes than Cohen (1988) through his books on power analysis. In these books, Cohen suggests general guidelines for levels of effect size. These are .2 for small effect, .5 for medium effect, and .8 for large effect. However, even Cohen was concerned about proposing these as standards. He stated:

> The terms "small," "medium," and "large" are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation. In the face of this relativity, there is a certain risk inherent in offering conventional operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioral science. This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference which is recommended for use only when no better basis for estimating the ES index is available. (1988, p. 25)

Cohen's concerns were cited by Wolf (1986) and suggests that effect sizes should be interpreted in context. Specifically, one possibility is to compare a given effect size to the median effect size of studies extracted from the professional literature in that specific context rather than use some arbitrary guideline. Wolf indicates that a .5 standard deviation improvement is often considered practically significant and that the general guidelines of the National Institute of Education's Joint Dissemination Review Panel require .33 effect size, but at times will accept .25 to establish educational significance.

A broader debate on the use of statistical significance testing emerged from Cohen's power analysis books and other works. Kaufman (1998) indicates that the "controversy about the use or misuse of statistical significance testing has been evident in the literature for the past 10 years and has become the major methodological issue of our generation" (p. 1). The debate has spawned at least two special issues of journals (*Research in the Schools*, McLean & Kaufman, 1998; *Journal of Experimental Education*, Thompson, 1993) and dozens of other articles. The editorial policies of journals have been changed by the debate (e.g., APA, 1994; Schafer, 1990, 1991; Thompson, 1996, 1997).

The debate has ranged from those who recommend the elimination of statistical significance testing (e.g., Carver, 1978, 1993; Nix & Barnette, 1998) to those who staunchly support it (e.g., Frick, 1996; Levin, 1993, 1998; McLean & Ernest, 1998). However, even those who defend statistical significance testing indicate that significant results should be accompanied by a measure of practical significance. The leading method of reporting practical significance is through the provision of an effect size estimate (Kirk, 1996; McLean & Ernest, 1998; Robinson & Levin, 1997; Thompson, 1996). Unfortunately, the criteria for judging the practical significance of results based on effect size has defaulted to the use of Cohen's (1988) guidelines that even Cohen has warned us about (1988, 1990). As Wolf (1986) noted, empirical standards for judging effect size are needed.

While other studies have suggested that reasonably large effect sizes might occur by chance (Barnette & McLean, 1999, November), no other studies could be found that used the relationship between know factors (such as sample size and number of groups) and effect size to predict effect size. If such a relationship can be verified, it would help researchers avoid the over-interpretation of effect sizes.

Methods

A Monte Carlo study was set up using number of samples from 2 to 10 and sample sizes from 5 to 100 in steps of 5. In each number of samples and sample size configuration, 10,000 one-way ANOVA replications were generated, using samples drawn from the unit normal distribution, were conducted for a total of 1,800,000 replications. Data were generated using a program written in double-precision QuickBASIC. Analysis of the raw data was conducted using several routines of SAS®. The accuracy of this approach has been established in several other studies (e.g., Barnette & McLean, 1999, November).

Patterns of observed ES values were examined for their relationships with number and size of samples. Using these relationships, a regression equation was developed to predict effect size from number of subject per group and number of groups. Tables and figures were developed to show the results.

Results

First, the accuracy of the Monte Carlo procedures used can be seen by inspecting Table 1. Table 1 shows the obtained p-values for each of the preset alpha-values for the 1.8 million replications. It was clear that ES was influenced by both number and size of samples. A regression-based trend analysis was conducted to determine equations that could be used to predict the mean chance-based ES for given number and size of samples. It was determined that a power-type function of the form $a\, n^{-b}$ was the best fit of the observed data. The regression equation produced R^2 values that were virtually 1. Keeping in mind that all of the data were produced with the means being equal for all groups in each model, the mean eta-squared values for each sample size/number of groups combinations are shown in Table 2. Scanning across the rows and down the columns illustrates the trends.

Table 3 shows the eta-squared values as a power-type function of the sample size for each number of groups. The equations for determining these values is also shown. The results are even clearer when depicted as graphs. Figures 1-6 show the results for 2, 3, 5, 8, and 10 groups respectively. In each case the near-perfect fit of the regression lines is evident.

Here are a few examples of how this could be used:

Situation 1: K= 2, n= 22
Observed Eta-squared= .1876
Predicted Eta-squared= .0235
Proportion of variance accounted for by treatment above what would be expected by chance (the value added)= .1641

Situation 2: K= 5, n= 50
Observed Eta-squared= .2215
Predicted Eta-squared= .0161
Proportion of variance accounted for by treatment above what would be expected by chance (the value added)= .2054

Situation 3: K= 8, n= 7
Observed Eta-squared= .1134
Predicted Eta-squared= .1268
Proportion of variance accounted for by treatment above what would be expected by chance (the value added)= 0

Discussion and Recommendations

It is obvious that one can use these results to estimate the eta-squared that might be expected by chance. In a given situation, subtracting the predicted chance eta-squared from the eta-squared obtained in an experiment would give the proportion of variance that could be attributed to the treatment beyond what would be expected by chance. Such an approach would remove the bias cited as the major weakness of the use of eta-squared as a measure of strength of association and make it a more useful measure of non-chance.

We recommend that these results be replicated and if proved to be valid, the use of the corrected eta-squared coefficient become common practice. At the very least, when an eta-squared value is cited, the chance eta-squared is presented for comparison. One limitation of this research is that equal sample sizes were used. For this procedure to have maximum utility, predicting the chance eta-squared when unequal samples sizes are used is needed.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Barnette, J. J. & McLean, J. E. (1999, November). Empirically based criteria for determining meaningful effect size. A paper presented at the 1999 Annual Meeting of the Mid-South Educational Research Association, Point Clear, AL.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312.

Fowler, R. L. (1988). Estimating the standardized mean difference in intervention studies. *Journal of Educational Statistics, 13*(4), 337-350.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*(4), 379-390.

Glass, G. V (1978). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5*, 351-379.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G. V, & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal, 6*, 403-414.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). *Meta-Analysis in social research*. Beverly Hills, CA: Sage Publications.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect sizes from a series of experiments involving paired comparisons. *Journal of Educational Statistics, 18*(3), 271-279.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 2*(2), 107-128.

Hedges, L. V. (1982). *Statistical methodology in meta-analysis*. (ERIC Document Reproduction Service No. ED 227 133).

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*(1), 61-85.

Huberty, C. J. (1994). A note on interpreting an R-squared value. *Journal of Educational and Behavioral Statistics, 19*, 351-356.

Huynh, C. L. (1989). *A unified approach to the estimation of effect size in meta-analysis*. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 306 248).

Kaufman, A. S. (1998). Introduction to the special issue on statistical significance testing. *Research in the Schools, 5*(2), 1.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics, 8*(2), 93-101.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education, 61*(4), 378-381.

Levin, J. R. (1998). What if there were no more bickering about statistical significance tests? *Research in the Schools, 5*(2), 43-53.

Lynch, K. B. (1987). The size of educational effects: An analysis of programs reviewed by the Joint Dissemination Review Panel. *Educational and Policy Analysis, 9*(1), 55-61.

McLean, J. E. (1983). *A meta-analysis approach to impact evaluation of adoptions*. Paper presented at the National Diffusion Network Regional Meeting, Memphis, TN. (ERIC Document Reproduction Service No. ED 242 744).

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*(2), 15-22.

McLean, J. E., & Kaufman, A. S. (Eds.). (1998). Statistical significance testing [Special Issue]. *Research in the Schools, 5*(2).

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*(2), 3-14.

Reichardt, C. S., & Gollob, H. F. (1987). Taking uncertainty into account when estimating effects. *New Directions for Program Evaluation, 35*, 7-22.

Robinson, D. L., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York: Irvington.

Rosenthal, R. (1978). Combining the results of independent studies. *Psychological Bulletin, 85*, 185-193.

Schafer, W. D. (1990). Interpreting statistical significance. *Measurement and Evaluation in Counseling and Development, 23*, 98-99.

Schafer, W. D. (1991). Power analysis in interpreting statistical significance. *Measurement and Evaluation in Counseling and Development, 23*, 98-99.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*(2), 115-129.

Snijders, T. A. B. (1996). What to do with the upward bias in R-squared: A comment on Huberty. *Journal of Educational and Behavioral Statistics, 21*(3), 283-298.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61*(4), 334-349.

Thomas, H. (1986). Effect size standard errors for the non-normal non-identically distributed case. *Journal of Educational Statistics, 11*(4), 293-303.

Thompson, B. (Guest Ed.). (1993). Statistical significance testing in contemporary practice [Special Issue]. *The Journal of Experimental Education, 61*(4).

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1996). Rejoinder: AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 26*(5), 29-32.

Wolf, F. M. (1986). *Meta-Analysis: Quantitative methods for research synthesis.* Beverly Hills, CA: Sage Publications.

Table 1. Summary Statistics for Monte Carlo Replications, n= 1,800,000

Mean Probability of F	.500488
Observed p for $\alpha = .25$	.249857
Observed p for $\alpha = .10$	.100076
Observed p for $\alpha = .05$	.050373
Observed p for $\alpha = .01$	.010158
Observed p for $\alpha = .001$	.001007
Observed p for $\alpha = .0001$	.000099

Table 2. Eta-Squared by Number of Samples and Sample Size

n	K= 2	K= 3	K= 4	K= 5	K= 6	K= 7	K= 8	K= 9	K= 10	Total
5	.112864	.143262	.157646	.167354	.172747	.176884	.178897	.182443	.183633	.163970
10	.051924	.068719	.076761	.082179	.084605	.086731	.088563	.090047	.090855	.080043
15	.033855	.045719	.050117	.053768	.055857	.057598	.058721	.059857	.060517	.052890
20	.024876	.034059	.037886	.040475	.041863	.043147	.044052	.044762	.045354	.039608
25	.020277	.027106	.030216	.032373	.033936	.034588	.035044	.035775	.036050	.031707
30	.017220	.022748	.025163	.026851	.027680	.028713	.029539	.029726	.030025	.026407
35	.014539	.019534	.021628	.023054	.023852	.024458	.025089	.025386	.025776	.022591
40	.012597	.016680	.018761	.019931	.021085	.021516	.021839	.022326	.022506	.019693
45	.011259	.014940	.016644	.017896	.018672	.019169	.019478	.019844	.020047	.017550
50	.010230	.013601	.015142	.016069	.016808	.017151	.017597	.017829	.017946	.015819
55	.009067	.012254	.013549	.014549	.015161	.015632	.016005	.016162	.016474	.014317
60	.008367	.011189	.012508	.013288	.013824	.014338	.014509	.014794	.014970	.013087
65	.007597	.010385	.011587	.012350	.012860	.013192	.013479	.013708	.013882	.012115
70	.007272	.009617	.010718	.011476	.011887	.012270	.012467	.012750	.012824	.011253
75	.006640	.008939	.010020	.010776	.011119	.011510	.011668	.011908	.011919	.010500
80	.006321	.008358	.009379	.010032	.010410	.010715	.010963	.011118	.011276	.009841
85	.005962	.007815	.008793	.009437	.009828	.010106	.010349	.010474	.010548	.009257
90	.005557	.007408	.008306	.008935	.009261	.009531	.009707	.009886	.010006	.008733
95	.005272	.007063	.007901	.008372	.008747	.009115	.009237	.009326	.009502	.008282
100	.005021	.006639	.007466	.008041	.008324	.008621	.008755	.008884	.009039	.007866
Total	.018836	.024802	.027510	.029360	.030426	.031249	.031798	.032350	.032658	.028777

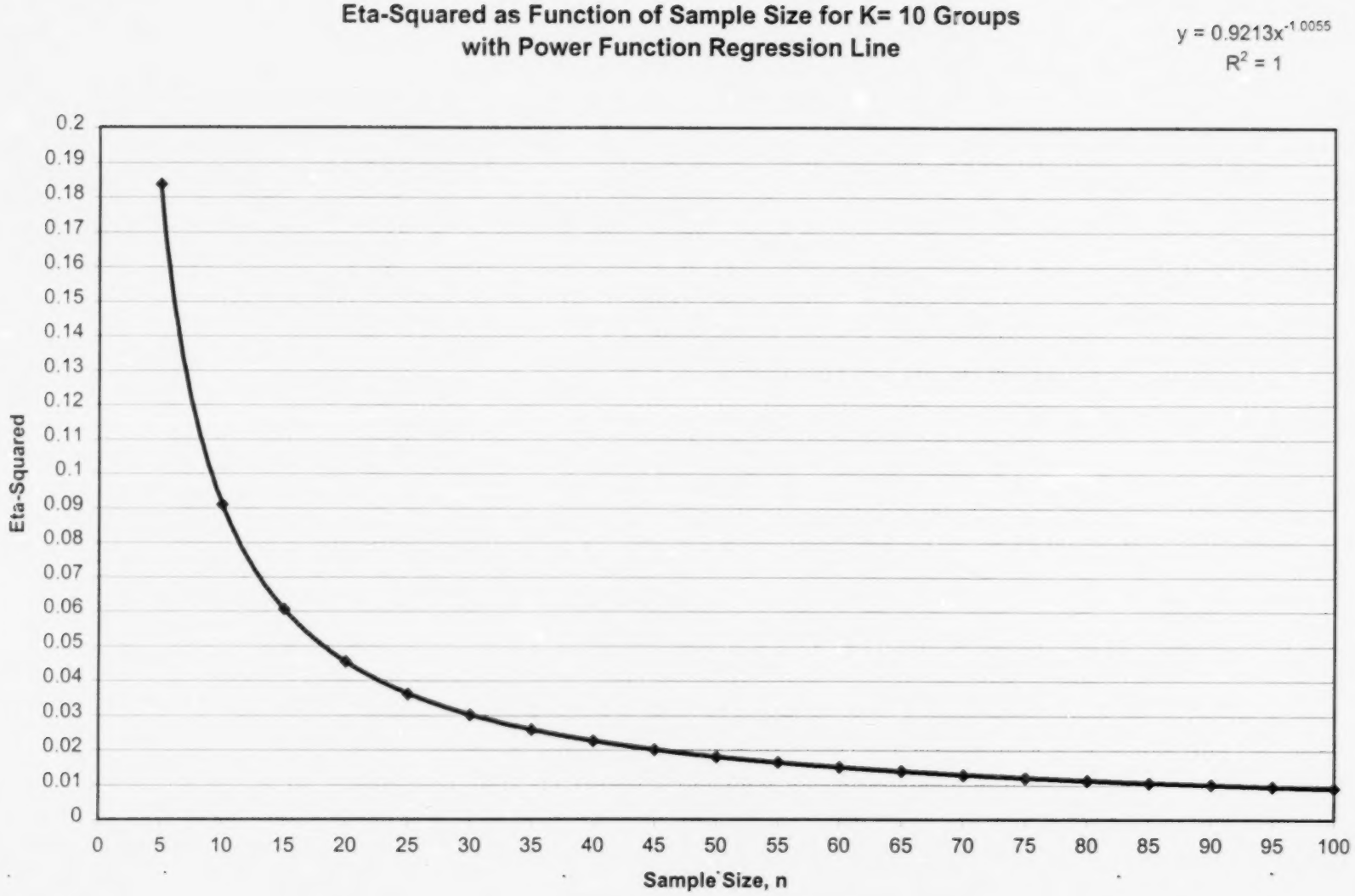
Table 3. Eta-Squared as a Function of Sample Size for Number of Groups, Eta-Squared = $a\ n^{-b}$

K	a	b	R^2
2	0.557324	1.024959	.999481
3	0.725375	1.018713	.999916
4	0.790802	1.012771	.999932
5	0.840826	1.011329	.999940
6	0.866752	1.009180	.999949
7	0.882383	1.006228	.999965
8	0.897923	1.005868	.999977
9	0.916229	1.06973	.999987
10	0.921270	1.005538	.999979

Coefficients a and b as function of K

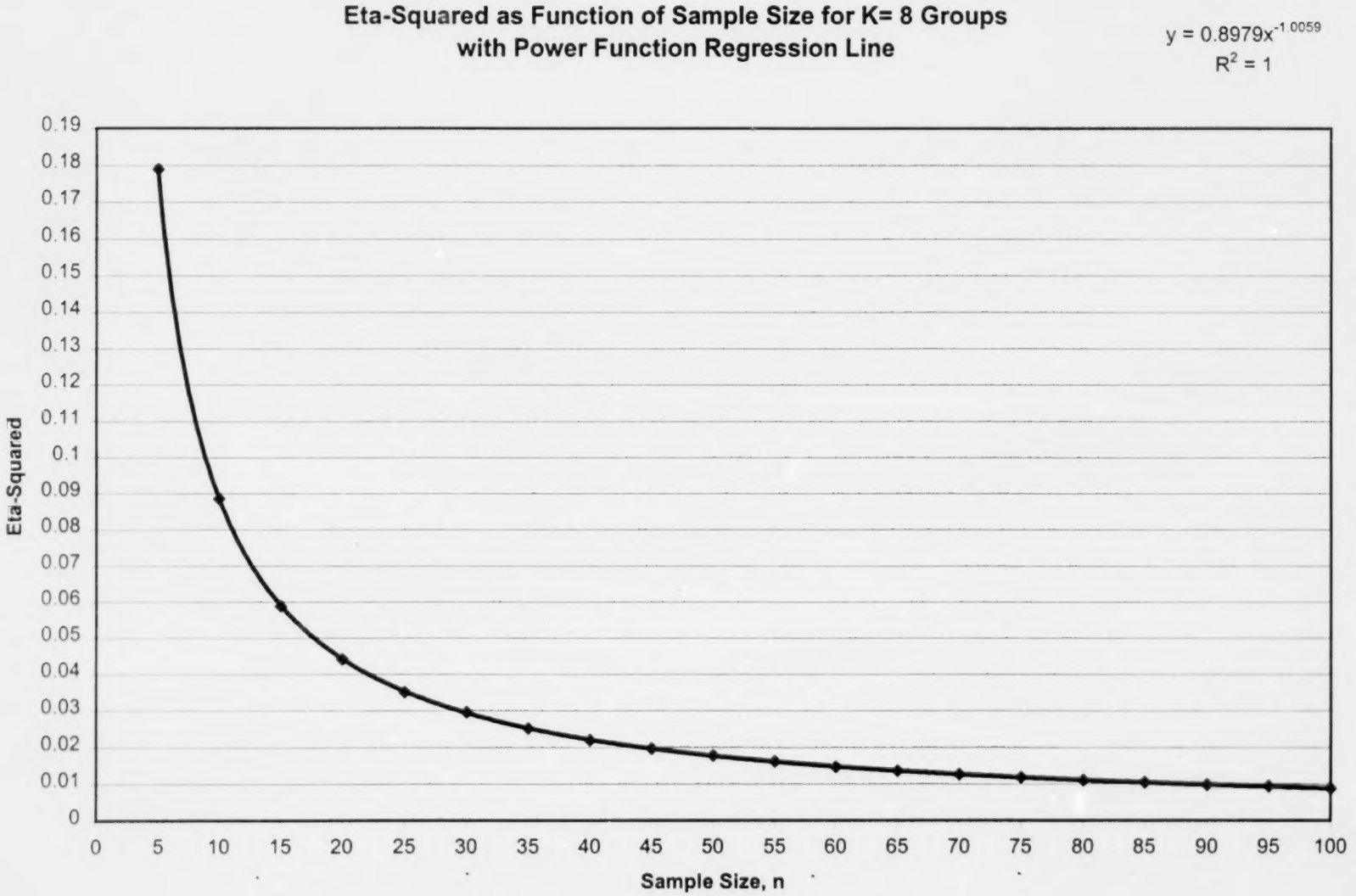
$a = 0.001677\ K^3 - 0.038036\ K^2 + 0.293553\ K + 0.120054$ ($R^2 = .991440$)

$b = -0.000046\ K^3 + 0.00255\ K^2 - 0.011803\ K + 1.043852$ ($R^2 = .987739$)


Eta-Squared as Function of Sample Size for K= 10 Groups
with Power Function Regression Line
y = 0.9213x^-1.0055
R^2 = 1
Eta-Squared
0.2
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Sample Size, n

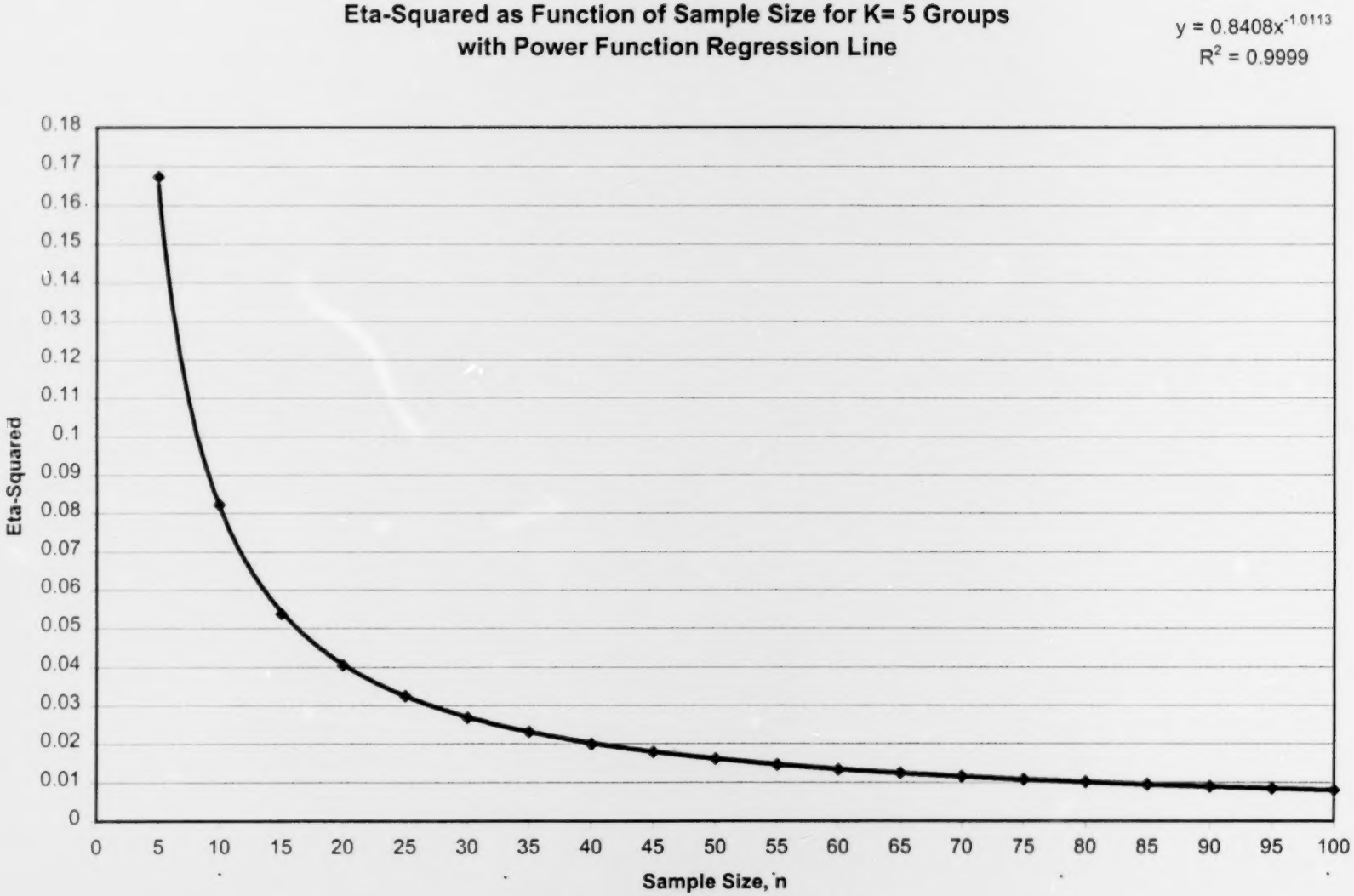
15

16


Eta-Squared as Function of Sample Size for K= 8 Groups
with Power Function Regression Line
y = 0.8979x^-1.0059
R^2 = 1
Eta-Squared
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Sample Size, n

17

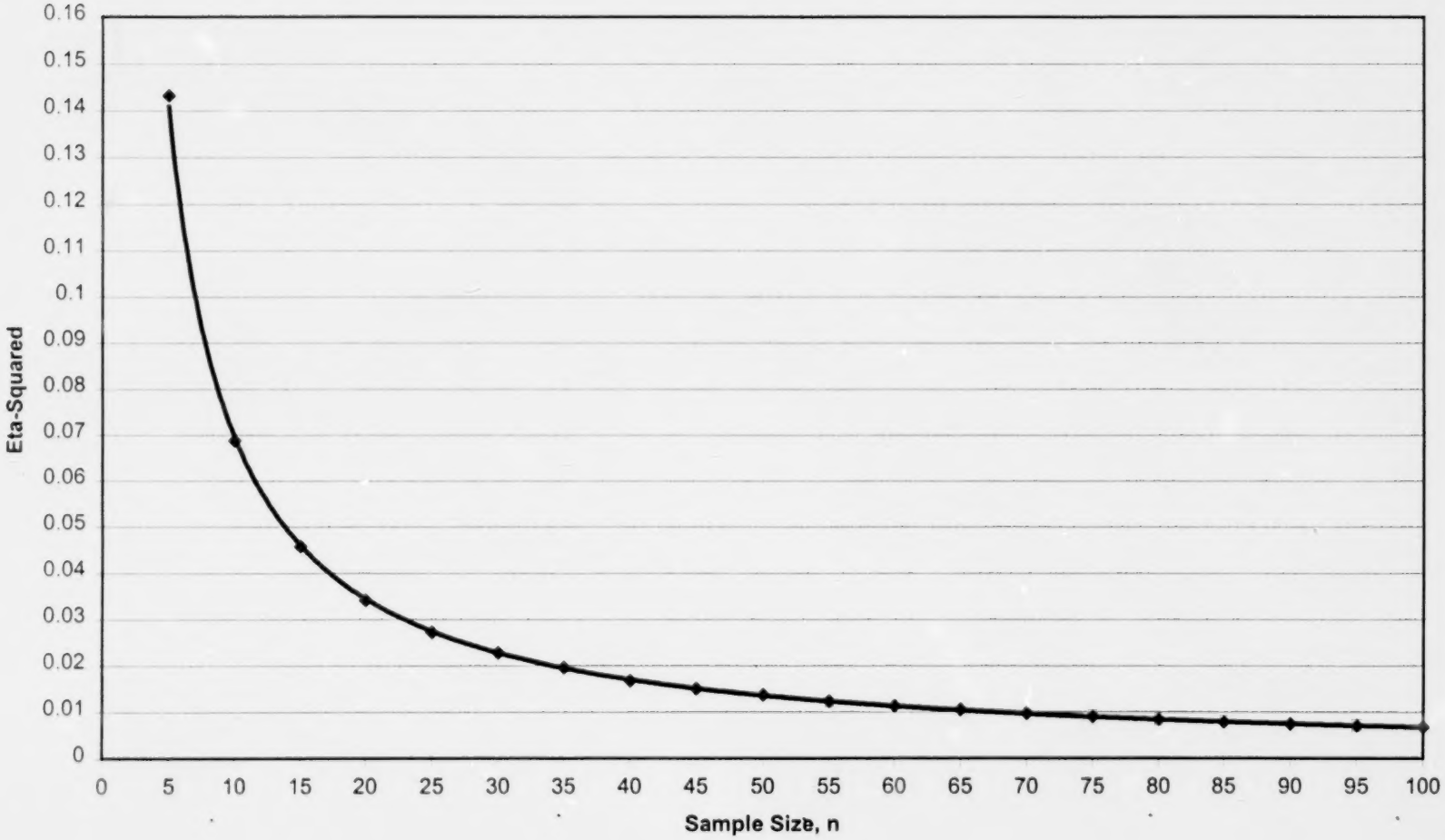
18


Eta-Squared as Function of Sample Size for K= 5 Groups
with Power Function Regression Line
y = 0.8408x^-1.0113
R² = 0.9999
Eta-Squared
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Sample Size, n

Eta-Squared as Function of Sample Size for K= 3 Groups with Power Function Regression Line

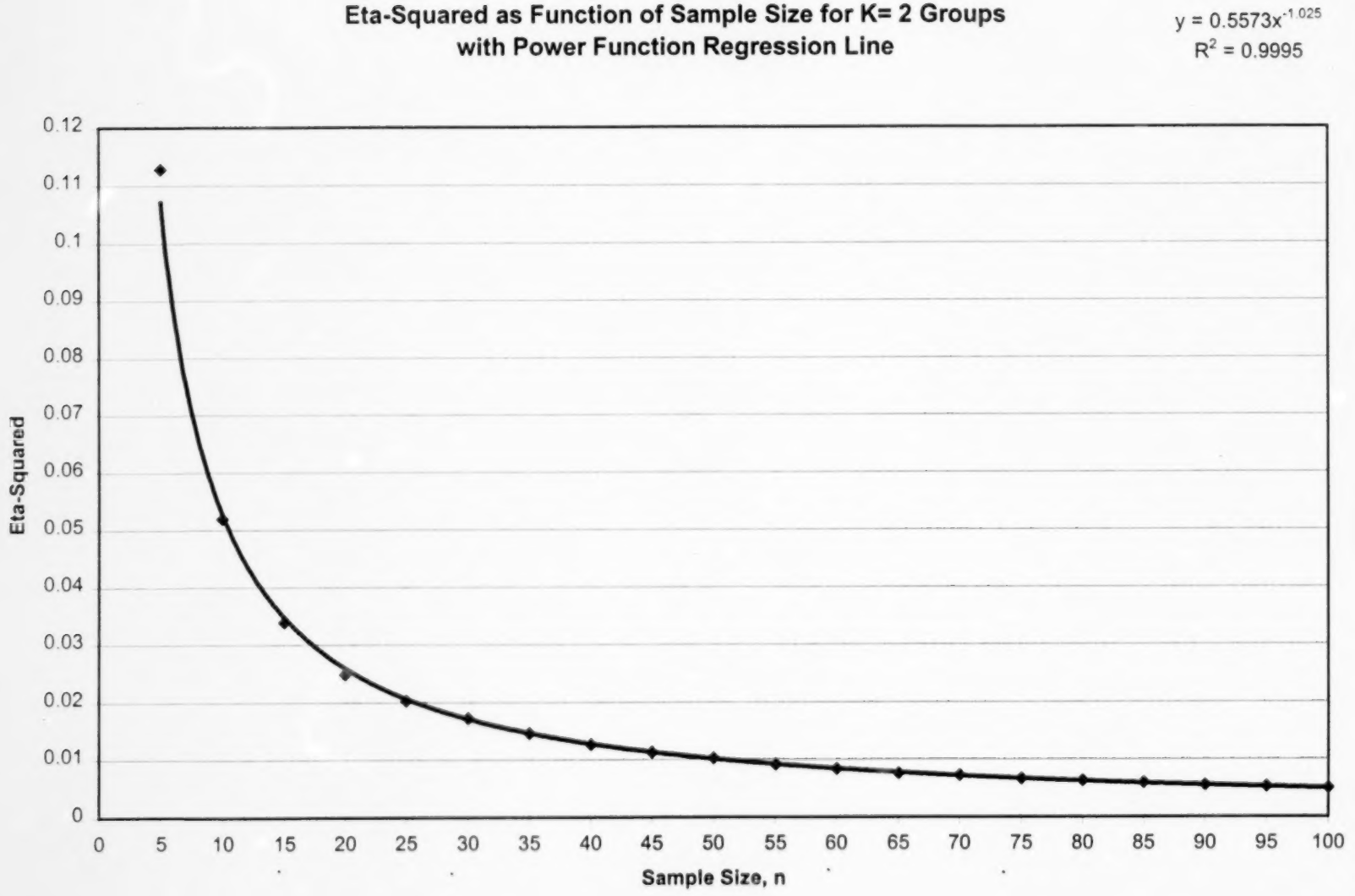
$y = 0.7254x^{-1.0187}$

$R^2 = 0.9999$



21

22


Eta-Squared as Function of Sample Size for K= 2 Groups
with Power Function Regression Line
y = 0.5573x^-1.025
R^2 = 0.9995
Eta-Squared
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Sample Size, n